Application of sequential pattern mining to the analysis of visitor trajectories

Nyoman Juniarta, Amedeo Napoli

Application of sequential pattern mining to the analysis of visitor trajectories

Nyoman Juniarta, Amedeo Napoli
nyoman.juniarta@loria.fr, amedeo.napoli@loria.fr

Introduction

In this work, we demonstrate the proof of concept of clustering 254 visitors based on their trajectories in a museum. We used a real dataset from Haifa Museum (Figure 1), where each trajectory is treated as a sequence of itemsets. We applied sim_{ACS} as a similarity measure between any two sequences.

START	END	ARTIFACT
…	…	…
12:47:23	12:47:29	GlassOvenVessels
12:48:34	12:48:39	WoodenTools
12:48:51	12:49:46	ReligionAndCult
…	…	…

Figure 1. An example of raw dataset from one visitor.

Behavior of visitors

According to [1], a visitor can be grouped as one of four defined behavior: ant, grasshopper, butterfly, and fish, as illustrated in Figure 2. These behaviors are described by the movement of a visitor around the artifacts in a museum.

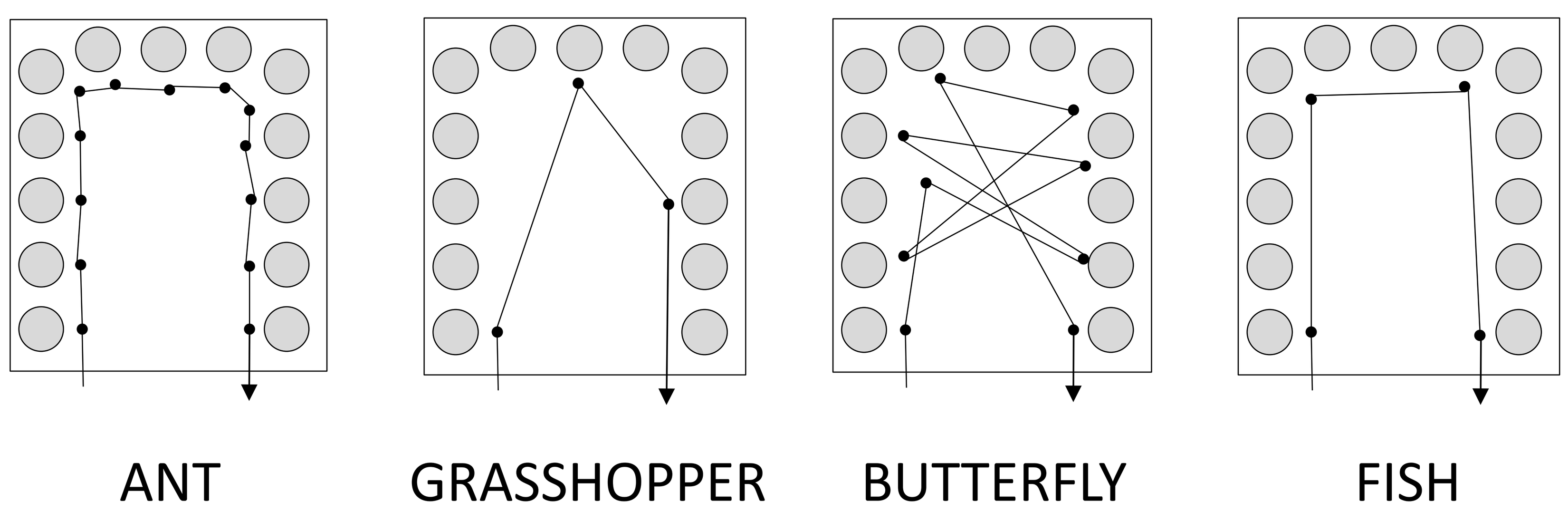

Figure 2. An example of raw dataset from one visitor.

Sequence

Sequence is an ordered list of itemsets. A sequence $s = \langle s_1, s_2, \dots s_m \rangle$ is a subsequence of $s' = \langle s_1, s_2, \dots s_n \rangle$ if there exist indices $1 \leq i_1 < i_2 < i_m \leq n$ such that $s_j \subseteq s'_{ij}$ for all $j = 1 \dots m$ and $m \leq n$.

$$S = \langle \{a, b\}, \{c, d\}, \{a, c, d\} \rangle$$

Subsequences	Not subsequences
$\langle \{a, b\}, \{d\} \rangle$	$\langle \{b,c\} \rangle$
$\langle \{c\}, \{a,c\} \rangle$	$\langle \{d\}, \{a,b\} \rangle$
…	…

Sequence clustering

In order to cluster a set of sequences, a similarity measure must be defined. Distance between any two sequences can be measured by sim_{ACS} [2], that counts all their common subsequences.

$$sim_{ACS} = \frac{\phi_A(s_1, s_2)}{\max\{\phi_D(s_1), \phi_D(s_2)\}}$$

$\phi_D(s)$ is the number of all distinct subsequences of s, and $\phi_A(s_1, s_2)$ is the number of all common subsequences between s_1 and s_2.

Clustering of visitors

In order to group visitors according to their trajectories, each trajectories was converted into a sequence of itemsets. An itemset corresponds to an artifact, and it possesses the information about *duration* and *distance* between previous artifact. These two elements were discretized such that *duration* has values *short* and *long*, while *distance* has *near* and *far*. Example:

$v_1 = \langle \dots, \{short, near\}, \{short, far\}, \{short, near\}, \dots \rangle$

$v_2 = \langle \dots, \{long, far\}, \{long, far\}, \{short, near\}, \dots \rangle$

…

Having the sequences of itemsets and non-Euclidean similarity measure, we applied hierarchical clustering over 254 sequences, using *hclust* from R software.

Result

Using 15 clusters as *hclust's* dendrogram cut, we obtained 4 big clusters and 11 small clusters, as shown in Figure 3.

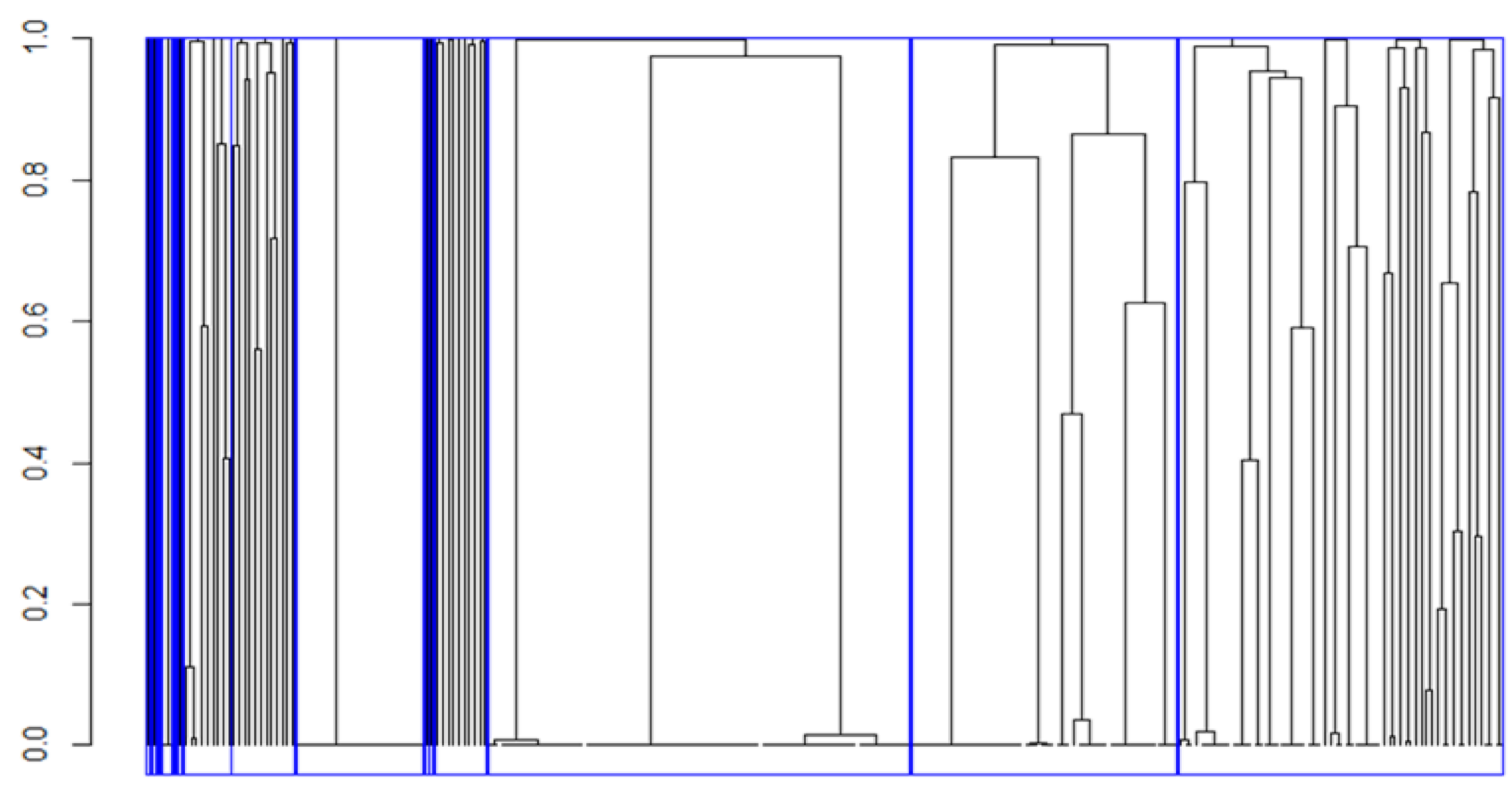

Figure 3. Result of hierarchical clustering over 254 visitors.

We calculated the presence of four possible itemsets in each cluster to map them into four visiting patterns. An *ant* corresponds to {*long*, *near*} itemset, while *fish* contains {*short*, *near*}. The {*long*, *far*} itemset can be correlated with both *butterfly* and *grasshopper*, but the latter has relatively short sequence.

Furthermore, within the 11 small clusters, the four possible itemsets have relatively the same support. This suggests that they correspond to visitors who frequently change their behavior while visiting the museum.

Acknowledgments

The thesis of Nyoman Juniarta is financed by the Région Grand-Est and the European project CrossCult.

References

[1] Massimo Zancarano et al. 2007. Analyzing museum visitors' behavior patterns. In *International Conference on User Modeling*. Springer, 238–246.

[2] Elias Egho et al. 2015. On measuring similarity for sequences of itemsets. *Data Mining and Knowledge Discovery* 29, 3, 732–764.